Introducing the mean absolute deviation 'effect' size

Stephen Gorard
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

This paper revisits the use of effect sizes in the analysis of experimental and similar results, and reminds readers of the relative advantages of the mean absolute deviation as a measure of variation, as opposed to the more complex standard deviation. The mean absolute deviation is is easier to use and understand, and more tolerant of extreme values. The paper then proposes the use of an easy to comprehend effect size based on the mean difference between treatment groups, divided by the mean absolute deviation of all scores. Using a simulation based on 1,656 randomised controlled trials each with 100 cases, and a before and after design, the paper shows that the substantive findings from any such trial would be the same whether raw-score differences, a traditional effect size like Cohen's d, or the mean absolute deviation effect size is used. The same would be true for any comparison, whether for a trial or a simpler cross-sectional design. It seems that there is a clear choice over which effect size to use. The main advantage in using raw scores as an outcome measure is that they are easy to comprehend. However, they might be misleading and so perhaps require more judgement to interpret than traditional 'effect' sizes. Among the advantages of using the mean absolute deviation effect size are its relative simplicity, everyday meaning, and the lack of distortion of extreme scores caused by the squaring involved in computing the standard deviation. Given that working with absolute values is no longer the barrier to computation that it apparently was before the advent of digital calculators, there is a clear place for the mean absolute deviation effect size (termed 'A').

Introduction and method

This paper introduces an 'effect' size calculation for a difference in mean scores between two groups, similar to Cohen's d, using the mean absolute deviation as the denominator. This is suggested as at least as robust and efficient, and easier to understand in many contexts, than an effect size based on the standard deviation. The paper starts by reminding readers of the role of standardised effect sizes, and continues to describe and compare the use of the mean absolute deviation and the standard deviation. The main part of the paper is based on a set of simulations, involving 1,656 pairs (64 multiples of the A-Z columns minus columns for labelling in Excel) of sets of random numbers between 0 and 1. Each pair is envisaged as being the before and after scores for 100 cases. Each pair yields a gain score, calculated as the value in the first column of the pair minus the second column (as in the example values for four cases in Table 1). The mean of each column is calculated, along with the mean absolute deviation and standard deviation for each gain score column. These values over 1,656 trials, and the difference between the means of the before and after scores, are then correlated with each other using Pearson's R. The paper ends by explaining how a mean absolute deviation effect size can be a useful analytical tool offering some of the advantages of both a standard deviation effect size, and a simple comparison of means.

Table 1 – Example of sub-set from one set of 100 random cases in one trial

Case	Pre-test score	Post-test score	Gain score	Deviation from mean gain score
1	0.507564	0.855072	-0.347510	0.320733
2	0.871944	0.636515	0.235429	0.262204
3	0.640988	0.913060	-0.272070	0.245297
4	0.239934	0.899633	-0.659700	0.632924
…	…	…	…	…
100	…	…	…	…
Mean	…	…	…	…

The role of effect sizes

There is a growth in the reporting of 'effect' sizes for numeric experimental and other empirical results for social science, as well as natural and health sciences. It is for example, the approach favoured over significance testing, or the citation of p-values, in the current publication manual of the American Psychological Association (2009). The US Institute of Education Science claim that significance testing and p-values are easily misunderstood, give misleading results about the substantive nature of results, and are 'best avoided' (Lipsey et al. 2012, p.3). Effect sizes are often needed in situations involving population data or non-random samples where p-values based on probabilistic uncertainty would be entirely inappropriate. Anyway, significance testing does not work as intended, and its use should cease (Gorard 2010). What is needed is instead a greater emphasis on straightforward reporting of results, within a clear research design, and placed in context so that the scale of the results and the size and quality of the dataset can be judged.

There is a number of different effect sizes used in conjunction with different types and distributions of data. These including shared variation, differences in variation, multiple groups, categorical variables and so on (Gorard 2013). This paper focuses on simple comparisons of means between two groups, assuming that the scores for both groups are for the same variable using the same scale of measurement. It is perfectly proper simply to report the two means and their difference in raw–score terms, especially where the scores already have a clear common meaning, or where they are already standardised (Lipsey et al. 2012). To say that a treatment group of young people improved their reading age by three months extra in comparison to a control is to state an 'effect' size, for example. Such a statement would also require a report of the sample size and quality, the quality of the measures used, and a reassurance of the comparativeness of the control group scores at the outset of the experiment.

In other circumstances, it is also reasonable to convert the results into a standardised effect size. This might be done in order to help readers understand the substantive importance of the result, or to allow the result to be synthesised with results from other studies perhaps using slightly different measures. It can be done with data from any comparative design, such as comparing the responses of two sub-groups in a survey, or presenting changes over time in a longitudinal study. A common method of creating a standard effect size is to divide the difference between two means by their standard deviation. In theory the SD to use here is that for the whole population. In the more realistic situation where only sample figures are available then the sample SD can be used instead, but even this compromise is ambiguous

and the subject of much dispute. If the experiment has a pre-test then the SD of the pre-test scores for both groups uses the largest number of cases that are unaffected by the experimental intervention. However, there is an inevitable delay with maturation between the pre-test and the post-test. Also the pre-test is rarely exactly the same as the post-test, in order to prevent practice effects. Both of these factors mean that the SD of the sample at pre-test may be a poor estimate of the SD of the population at post-test. Another possibility is to use the SD of the control group post-test scores. These are similarly unaffected by the intervention, presumably, and are more relevant to consideration of the outcome effect size. Unfortunately, the number of cases will inevitably be smaller than the combined total. So perhaps the best estimate of the SD will come either from the SD of the overall post-test scores, or the pooled SDs of the treatment and control group post-test scores. Given these and other variations, the standard effect size of difference between means divided by their standard deviation is not really that 'standard', with Cohen's d, Glass's delta, and Hedges' g and others all giving similar but different final results from the same datasets.

A further confusion is that many authorities argue that any such effect size should be accompanied by a confidence interval (often 95% probability). Confidence intervals are perhaps the most widely misunderstood basic component of statistics. They use the same underlying logic as significance tests and p-values (Coe 2002), and so suffer from the same limitations. They are irrelevant when dealing with population data, non-random samples and incomplete random samples (i.e. almost all real-life situations). And even in ideal conditions of distribution and sampling/allocation they provide a conditional probability that no analyst wants but which is usually misinterpreted as another conditional probability that is wanted. A confidence interval tells the reader, assuming that the population and sample effect sizes are identical, a range of values within which the effect sizes estimates of 95% of imaginary repeated samples of the same size would lie. Why anyone would want to know this is unclear. The confidence interval tells the reader nothing about how likely it is that the population and sample effect sizes are the same or even close (Gorard 2014a). This is a shame as this is what analysts want, and often imagine or pretend that they have. In fact, confidence intervals are useless for most real-life datasets, and have proven to be terribly misleading even when used with random samples.

The mean absolute deviation

As illustrated above, most standardised 'effect' size calculations involve a version of the standard deviation of the outcome measure, as a denominator. The standard deviation is a measure of 'dispersion' or 'spread' used as a common summary of the range of scores associated with a measure of central tendency – the mean-average. It is obtained by summing the squared values of the deviation of each observation from the mean, dividing by the total number of observations, and then taking the positive square root of the result. Although the average of the squared deviations from the mean is then square-rooted, this reversal is incomplete and so SD tends to overemphasise the more extreme deviations.

An alternative summary of dispersion is the mean absolute deviation (M|D|). This is simply the mean average of the absolute differences between each score and the overall mean - the amount by which, on average, any figure differs from the overall mean. The mean absolute deviation gives each deviation its proportionate place in the result, and is easier for new researchers and others to understand than SD is. New researchers struggle with the logic of statistics (Watts, 1991, Murtonen and Lehtinen 2003), especially measures of variation

(Cooper and Shore 2008), perhaps because it is made unnecessarily complicated by experts and practioners. Working with empirical astronomical data, Eddington (1914, p.147) found, and reported that the majority of astronomers also found, that the mean absolute deviation was a better measure of dispersion than SD.

The widespread use of SD stems from the writings of Fisher (1920), who argued that SD was more 'efficient' than M|D| under ideal circumstances (i.e. not for the pragmatic context cited by Eddington). SD and M|D| can be considered equivalent in terms of consistency and sufficiency. M|D| is calculated in the same way, whether the calculation is for a true random sample or for the known population from which that sample is drawn. The same is true for SD. They are both consistent. Similarly, when working with a true random sample and a known population, both SD and M|D| can be used to summarise all of the relevant information to be gleaned from the sample about the population parameter. Both are 'sufficient'. It is therefore, mainly in terms of efficiency that SD and M|D| may differ in quality. This means that when, and only when, working with a true random sample and a population, the chosen statistic (SD or M|D|) for the sample should have the smallest probable error as an estimate of the equivalent population parameter.

When drawing repeated large samples from a normally distributed population, the standard deviation of their individual mean deviations is 14% higher than the standard deviation of their individual standard deviations (Stigler 1973). Thus, the SD of such a sample is a more consistent estimate of the SD for a population, and is considered better than its plausible alternatives as a way of estimating the standard deviation in a population using measurements from a sample (Hinton 1995, p.50). This is the basis of Fisher's argument, and a major explanation for why SD has been preferred to M|D|, and so why much of subsequent statistical theory is based on it.

Unfortunately, this argument does not stand up to even superficial scrutiny. Fisher (1920) assumed as a premise for the argument that which he was supposedly trying to establish. It does not matter that the standard deviation of the individual mean deviations of repeated large samples drawn from a normally distributed population is higher than the standard deviation of their individual standard deviations. This is what one would expect. In the same way, the mean deviation of the individual standard deviations for repeated random samples is higher than the mean deviation of their individual mean deviations. It is crucial to compare like with like. The mean deviation of a sample is a better estimate of the mean deviation of a population than the standard deviation is, and *vice versa*. It is also the case that efficiency is irrelevant in the most common situations in social science where an analyst wishes to describe variation among their cases but does not have a true random sample, or is working with population data anyway and so the issue does not arise.

Fisher's argument, and the evidence for it, is anyway premised on working with data that is strictly normally distributed in the population. It was never applied to the more realistic situation of dealing with observations that merely approximate such an ideal. If the data is not known to be normally distributed then the whole argument is irrelevant, and M|D| has been demonstrated to be better for use with distributions other than the normal distribution (Stigler 1973). The calculation of the relative efficiency of SD and M|D| also depends on there being no errors at all in the observations (Tukey 1960). But even for normal distributions with tiny errors in the data, which could include a few missing cases as well as measurement errors, the superior efficiency of the M|D| is evident (according to Barnett and Lewis 1978, p.159, Huber 1981, p.3). So even using Fisher's faulty logic, M|D| is still actually more efficient in all life-

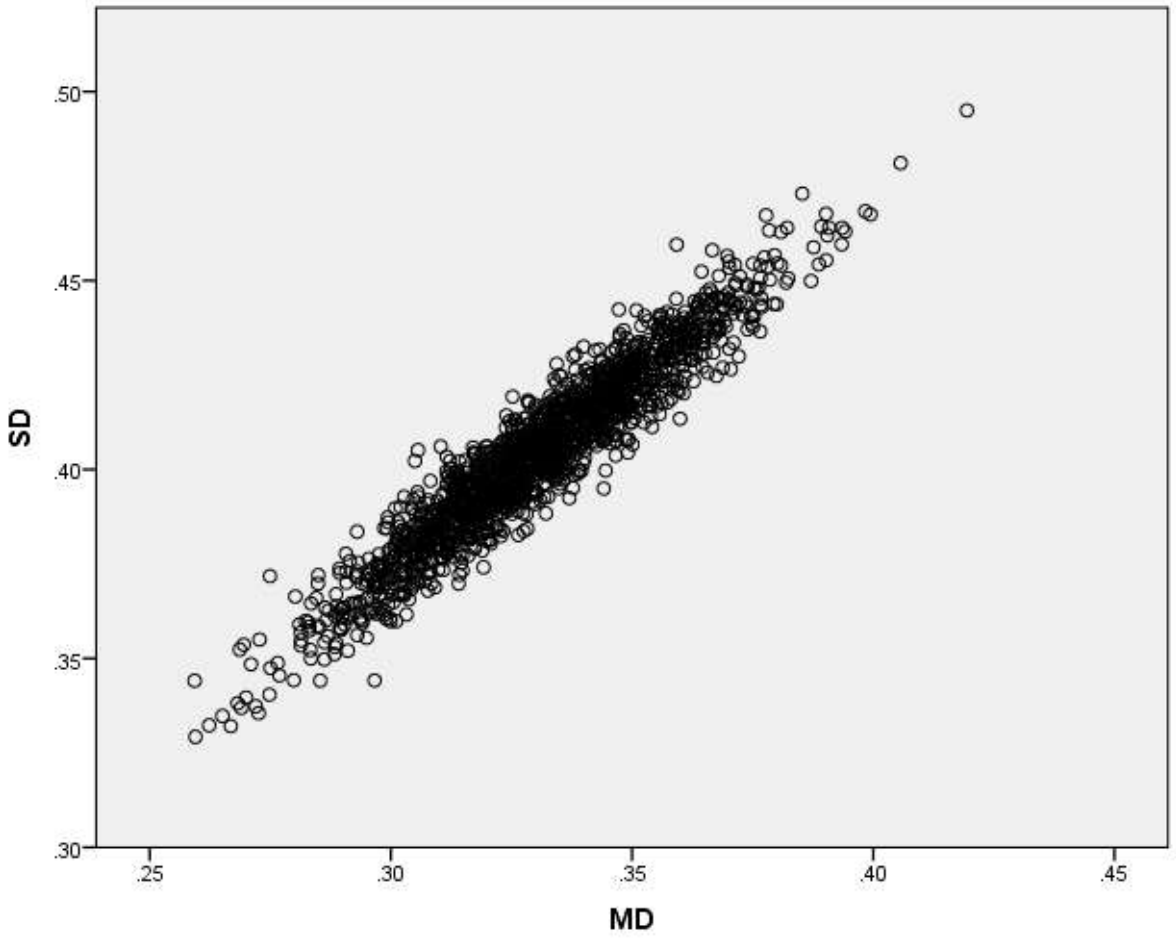
like situations where small errors will occur in observation and measurement. Use of M|D| can 'ensure high stability of statistical inference when we deal with distributions that are not symmetric and for which the normal distribution is not an appropriate approximation' (Amir 2012, p.145).

Repeated simulations show that the efficiency of M|D| is at least as good as SD, here illustrated for uniform distributions. For example, the population of 20 integers between 0 and 19 has a mean of 9.5, a mean absolute deviation of 5, and a standard deviation of 5.77. This is a clear example of how M|D| gives an easy to interpret and predictable answer that more readers can understand than the SD. Run as a simulation a set of 1000 samples (with replacement) of 10 random numbers - each between 0 and 19 – had sample SDs varying from 2.72 to 7.07, and sample M|D| s varying from 2.30 to 6.48. The standard deviation of the 1,000 estimated standard deviations around their true mean of 5.77 was just over 1.025. The standard deviation of the 1,000 estimated mean deviations around their true mean of 5 was just under 1.020. These values and their direction of difference are relatively stable over repeated simulations with further sets of 1,000 samples. This is an illustration that, for uniform distributions of the kind involving random numbers, the efficiency of the mean deviation is actually better than that of the standard deviation (Gorard 2005).

Real-life datasets of social scientific measurements tend to have distributions with more extreme scores than would be expected under Fisher's ideal assumptions. This means that squaring the deviations from average to produce the SD increases the apparent variation, and the act of square-rooting the sum of squares does not completely reverse this (Huber 1981). The distortion caused by this squaring then forms part of the pressure on analysts to ignore any extreme values or 'outliers' (Barnett and Lewis 1978). Yet extreme scores are important valid occurrences in a variety of natural phenomena such as city growth, income distribution, earthquakes, traffic jams, solar flares, and avalanches. For these, statistical techniques based on the mean absolute deviation are preferred (Fama 1963, p.491).

The issue of whether to work with M|D| or SD in conducting any analysis is important, because the results obtained will differ. Figure 1 illustrates this, and is based on analysis of 1,656 sets of 100 random numbers between 0 and 1. It shows that for the same data, the standard deviation (SD) and mean absolute (M|D|) are closely related. SD is always larger than M|D| because the initial squaring of deviations in the calculation of SD tends to exaggerate them. Their Pearson R correlation over the 1,656 trials represented here was +0.953. This means that less than 91% of the variation is common to both measures of dispersion. The choice is therefore not like deciding whether to use imperial or metric measurements, for example, where there is a direct conversion from one to the other. The 'width' of the scatter in Figure 1 clearly shows that any value calculated for M|D| will have more than one equivalent SD (and vice versa). Over 9% of the variation is not common, and that would be a large amount to be unaccounted for in most analyses. Some commentators ~~authorities~~ ~~(e.g.~~ http://forumserver.twoplustwo.com/47/science-math-philosophy/formula-quickly-convert-standard-deviation-absolute-mean-deviation-1195849/) ~~)~~ quote a 'simple' conversion between the two as though they were linearly related: M|D| =SD.√(2/p), which is approximately the same as M|D| = 0.79788456 times SD. Hoaglin et al. (1983) suggests that the standard deviation is approximately 1.4826 times the absolute value. However, ~~this~~ either value applies only to a perfect normal distribution, of the kind never seen in real research. In this simulation, for example, the mean of the SDs was 0.405 and the mean of the M|D| s was 0.331. Therefore, SD=M|D| times 0.331/0.405 or M|D| times 0.81728395. The simple conversion does not work, because there is a real difference between the two measures.

Figure 1 – Scatterplot of equivalent SD and M|D| for 1,656 simulated experiments



Note: The figure of 1,656 was a convenient multiple (64 minus 18 columns for labelling) when doubling up the alphabetic columns in Excel. It was no significance.

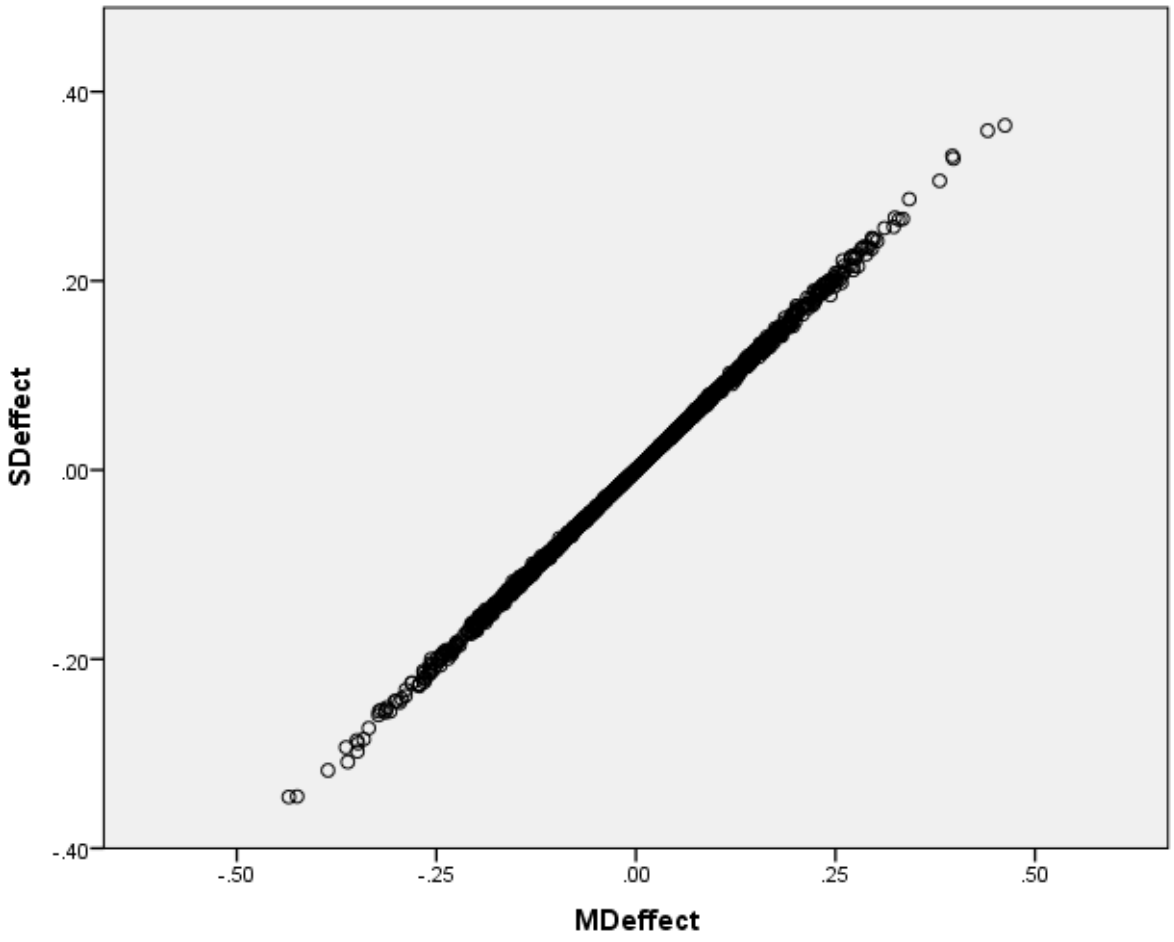
All of this has led to the mean absolute deviation being used routinely in a number of areas other than the astronomy of Eddington, including biology, engineering, IT, physics, imaging, geography and environmental science (e.g. Anand and Narasimha 2013, Hao et al. 2012, Hižak J. and Logožar R. 2011, Sari et al. 2012). In each case, M|D| is preferred for its ease of understanding, unbiased treatment of all scores (whether extreme or not), accuracy or efficiency, or simply because the results are found to be easier to portray. Oddly, despite all of the above papers citing a methods paper published in the social sciences (Gorard 2005), social sciences is one area where this mini-revolution has not yet picked up. For many authors the use of absolute numbers is no longer a barrier to computation. M|D| is linked to a range of other simple analytical techniques, again with relatively easy to understand meanings, largely because they do not require the squaring and square rooting of differences (Gorard 2005). These include a 'segregation index' (GS) for summarising the unevenness in the distribution of individuals between organisational units (Gorard and Taylor 2002), the relative difference, or achievement gap (Gorard et al. 2001), and an absolute deviation correlation coefficient (Gorard 2014b).

The mean absolute deviation effect size

The choice of whether the M|D| or SD is most suitable as a measure of dispersion then relates to which should be used in calculating an 'effect' size. Mainstream argument about which effect size to use has been focused on which standard deviation is most appropriate for the denominator – that of the pre-test, control group, pooled groups etc. But, however this issue is resolved, another clear alternative would be to use the mean absolute deviation to create a mean absolute deviation effect size (hereafter referred to as 'A' for brevity). For a simple experimental design, 'A' would be the difference between the mean outcome scores for both groups divided by the mean absolute deviation of the scores (for the pre-test, control group, or pooled groups etc.). For the examples that follow 'A' is based on the mean absolute deviation of the gain scores from pre- to post-test in the repeatedly simulated trial.

Figure 2 shows the relationship between the effect sizes for the same 1,656 simulated trials as Figure 1, using both standard deviation and mean absolute deviation. It shows that the scatter in the relationship is less than for the standard deviation and mean absolute deviation themselves (Figure 1). As expected, since SD is always larger than the equivalent M|D|, the M|D| effect size is larger than the SD effect size for any set of data. However, except at the extremes, it seems to make very little difference which version of the effect size is used.

Figure 2 – Scatterplot of equivalent SD and M|D| 'Effect' sizes for 1,656 simulated experiments



This strong relationship is then reflected in the correlation coefficients in Table 2. Here the results of the 1,656 simulated trials are presented as a simple difference between mean scores for two treatment groups, and converted into two alternative 'effect' sizes, one based on the

standard deviation (d) and one on the mean deviation (Aa). To three decimal places, A and d are indistinguishable. The two measures of dispersion themselves are unrelated to the effect sizes and the difference between means is unrelated to the two measures of dispersion. As noted at the start of the paper, the simple raw difference between means is a perfectly accurate 'effect' size, which correlates almost perfectly with both the M|D| and SD effect sizes. With data such as these, it really does not matter which of the three effect sizes or which of the measures of dispersion is used. They will presumably yield the same substantive conclusion in practice, except where the true underlying effect size has a very large absolute value.

Table 2 – Correlation between five experimental outcomes

	Difference between means	SD 'Effect' size	M|D| 'Effect' size (A)	Standard deviation	Mean |deviation|
Difference between means		+.998	+.997	+.001	-.007
SD 'Effect' size (d)	+.998		1	-.005	-.013
M|D| 'Effect' size (a)	+.997	1		-.006	-.014
Standard deviation	+.001	-.005	-.006		+.953
Mean |deviation|	-.007	-.013	-.014	+.953	

N=1,656

Discussion

As a measure of dispersion and as the denominator for calculating effect sizes, the standard deviation has one chief advantage. It is already in widespread use, and has been for at least 100 years since the dispute between Eddington and Fisher. By definition it is linked to the normal distribution which also means that it appears in many statistical settings and guises. This is an important factor when selecting a standard effect size. However, it is hard to teach to new researchers, and has no easy to understand meaning in real-life. It also exaggerates the importance of extreme scores for no clear reason, promoting the deletion of purported 'outliers', and is less efficient than the mean absolute deviation in the realistic situation where data is not in an ideal normal distribution, or where it has any errors at all.

Using raw scores instead, such as the mean difference in outcomes between two experimental groups, is superficially simple. It seems easy to explain to research users, and for new researchers to learn about. And as the analysis in Table 2 shows, it can yield an equivalent result to 'effect' sizes based on a measure of dispersion (where the scores are uniform, well-understood, or already standardized). Eliminating the use of p-values and confidence intervals already makes data analysis from robust designs like true experiments much easier to conduct and to report to users. This is an under-rated but huge advantage for explanation, critique and for capacity-building. Eliminating something as intrinsically strange for newcomers as the standard deviation, as well, would be a further boost to the security and the openness of research results. However, there are many cases where simple differences, or

even differences in proportions, have been misinterpreted (Gorard 1999). Using the simple difference approach requires a little more care in judgement about the research design and the context for the scores than appears at first sight.

Therefore, the mean absolute deviation effect size, A, is presented here as a form of compromise. It can be used wherever an 'effect' size is appropriate. It is much easier than the SD effect size for everyone to understand and with an everyday meaning (Gorard 2006). It at least as safe as the SD effect size, and requires a little less care to work with than raw scores alone. Like the SD, the mean absolute deviation is already in use in a variety of fields and that use is growing. The links with wider statistical and analytical issues will also grow if the stranglehold of the standard deviation is further challenged, instead of placing our 'reliance in practice on isolated pieces of mathematical theory proved under unwarranted assumptions, [rather] than on empirical facts and observations' (Hampel 1997, p.9). A new form of statistics with a mean absolute deviation effect size, mean absolute deviation correlation, least absolute deviation regression models and so on is possible. In most contexts it would be more robust (e.g. Amir 2012), and could fit together better (e.g. Cahan and Gamliel 2011). The purported mathematical advantage of the SD under ideal circumstances is based on an error of logic, does not apply to datasets with any imperfections, and is irrelevant when not dealing with a complete random sample (or allocation). The computational problems with absolute numbers have been effectively solved by the power of modern computing. The main remaining drawback to the use of a metric with absolute numbers in it therefore concerns algebraic manipulation. However, most potential analysts do not want to carry out any algebraic manipulation. Most might therefore appreciate descriptive statistics, whether for populations, samples or simply groups of cases, which are more democratic than currently by being easier to compute and easier to understand.

References

American Psychological Association (2009) *Publication manual of the APA 6th Edition*, Washington DC: American Psychological Association

Amir, E. (2012) On uses of mean absolute deviation: decomposition, skewness and correlation coefficients, *METRON – International Journal of Statistics*, LXX, 2-3, 145-164

Anand, M. and Narasimha, Y. (2013) Removal of salt and pepper Noise from highly corrupted images using mean deviation statistical parameter, *International Journal on Computer Science and Engineering*, 5, 2, 113-119

Barnett, V. and Lewis, T. (1978) *Outliers in statistical data*, Chichester: John Wiley and Sons

Cahan, S. and Gamliel, E. (2011) First among others? Cohen's d vs. alternative standardized mean group difference measures, *Practical Assessment, Research and Evaluation*, 16, 10, 1-6

Coe, R. (2002) It*'s the effect size, stupid*, presentation to British Annual Research Association Conference, Exeter, September 2002

Cooper L. and Shore F. (2008) Students' Misconceptions in Interpreting Center and Variability of Data Represented via Histograms and Stem-and-leaf Plots, *Journal of Statistics Education*, 16, 2, 13 pages

Eddington, A. (1914) *Stellar movements and the structure of the universe*, London: Macmillan

Fama, E. (1963) Mandelbrot and the stable Paretian hypothesis, *Journal of Business*, pp. 420-429

Fisher, R. (1920) A mathematical examination of the methods of determining the accuracy of observation by the mean error and the mean square error, *Monthly Notes of the Royal Astronomical Society*, 80, 758-770

Gorard, S. (1999) Keeping a sense of proportion: the "politician's error" in analysing school outcomes, *British Journal of Educational Studies* , 47, 3, 235-246

Gorard, S. (2005) Revisiting a 90-year-old debate: the advantages of the mean deviation, *The British Journal of Educational Studies*, 53, 4, 417-430

Gorard, S. (2006) *Using everyday numbers effectively in research:* Not *a book about statistics*, London: Continuum

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, 36, 1, 63-77

Gorard, S. (2013) *Research Design: Robust approaches for the social sciences*, London: SAGE

Gorard, S. (2014a) The widespread abuse of statistics by researchers: what is the problem and what is the ethical way forward?, *Psychology of Education Review*, 38, 1, 3-10

Gorard, S. (2014b) A 'new' correlation coefficient, https://www.dur.ac.uk/education/research/current_research/quantitative-methods/resources/, accessed 25/3/14

Gorard, S. and Taylor, C. (2002) What is segregation? A comparison of measures in terms of strong and weak compositional invariance, *Sociology*, 36, 4, 875-895

Gorard, S., Rees, G. and Salisbury, J. (2001) The differential attainment of boys and girls at school: investigating the patterns and their determinants, *British Educational Research Journal*, 27, 2, 125-139

Hampel, F. (1997) *Is statistics too difficult?*, Research Report 81, Seminar fur Statistik, Eidgenossiche Technische Hochschule, Switzerland

Hao, Y., Flowers, H., Monti, M. and Qualters, J. (2012) U.S. census unit population exposures to ambient air pollutants, *International Journal of Health Geographics* 2012, 11, 3 doi:10.1186/1476-072X-11-3

Hinton., P. (1995) *Statistics explained*, London: Routledge

Hižak J. , Logožar R. (2011) A derivation of the mean absolute distance in one-dimensional random walk, *Technical Journal*, 5, 1, 10-16

Hoaglin, D., Mosteller, F. and Tukey, J. (1983) *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons

Huber, P. (1981) *Robust Statistics*, New York: John Wiley and Sons

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., Roberts, M., Anthony, K. and Busick, M. (2012) *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*, Washington DC: Institute of Education Sciences

Murtonen, M. and Lehtinen, E. (2003) Difficulties experienced by education and sociology students in quantitative methods courses, *Studies in Higher Education*, 28, 2, 171-185

Sari, S., Roslan, H. and Shimamura, T. (2012) *Noise estimation by utilizing mean deviation of smooth region in noisy image*, Fourth International Conference on Computational Intelligence, Modelling and Simulation, http://eprints.uthm.edu.my/3265/1/4871a232.pdf

Stigler, S. (1973) Studies in the history of probability and statistics XXXII: Laplace, Fisher and the discovery of the concept of sufficiency, *Biometrika*, 60, 3, 439-445

Tukey, J. (1960) A survey of sampling from contaminated distributions, in Olkin, I., Ghurye, S. , Hoeffding, W., Madow, W and Mann, H. (Eds.) *Contributions to probability and statistics: essays in honor of Harold Hotelling*, Stanford: Stanford University Press

Watts, D. (1991) Why is introductory statistics difficult to learn?, *The American Statistician*, 45, 4, 290-291